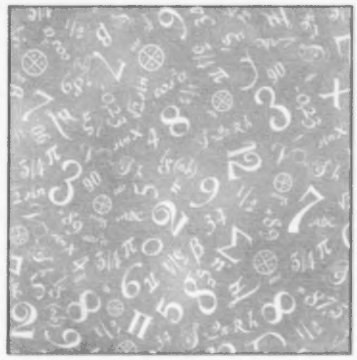
Composante du produit nº 82-003-X au catalogue de Statistique Canada
Rapports sur la santé

Article

Stratégies de traitement des hypothèses de normalité dans la modélisation multiniveaux : une étude de cas d'estimation des trajectoires des scores du Health Utilities Index Mark 3



par Julie Bernier, Yan Feng et Keiko Asakawa

Décembre 2011

Statistique Canada Statistics Canada

Canada

Stratégies de traitement des hypothèses de normalité dans la modélisation multiniveaux : une étude de cas d'estimation des trajectoires des scores du Health Utilities Index Mark 3

par Julie Bernier, Yan Feng et Keiko Asakawa

Résumé

Contexte

Les données longitudinales permettent d'étudier la dynamique de l'état de santé au cours du cycle de vie en modélisant les trajectoires. Les trajectoires de l'état de santé mesurées au moyen du *Health Utilities Index Mark 3* (HUI3) modélisées sous forme d'une fonction de l'âge seulement, ainsi que d'une fonction de l'âge et de covariables socioéconomiques, ont révélé des résidus non normaux et des problèmes d'estimation de la variance. Le but de l'étude était d'examiner la possibilité de transformer la distribution des scores HUI3 de manière à obtenir des résidus qui suivent approximativement une loi normale.

Données et méthodes

L'analyse est fondée sur des données longitudinales provenant des six premiers cycles de l'Enquête nationale sur la santé de la population (ENSP). Les données ont été recueillies auprès de 7 784 personnes qui, en 1994-1995, étaient âgées de 40 à 99 ans, vivaient dans des ménages privés et avaient fourni des renseignements complets pour le calcul de l'HUI3. Un modèle de croissance multiniveaux a été utilisé pour examiner la structure hiérarchique des données de l'ENSP (mesures répétées emboîtées dans les enquêtés). La transformation arcsin[2 × (HUI + 0,36) / (1 + 0,36) - 1] a été appliquée pour améliorer la distribution des résidus aux deux niveaux et pour limiter la moyenne conditionnelle à l'intervalle de -0,36 à 1,00. Un modèle a été estimé en utilisant les déterminants socioéconomiques de la santé. Les analyses ont été effectuées à l'aide des logiciels SAS et MLwiN.

Résultats

Après la transformation des scores HUI3, le modèle était satisfaisant et permettait d'ajouter de nouvelles variables sociodémographiques et d'état de santé afin d'estimer leur effet sur la qualité de vie liée à la santé de populations vieillissantes. Étant donné la transformation complexe du modèle arcsinus, les coefficients de régression n'ont pas été interprétés. Les résultats des estimations ont plutôt été résumés graphiquement.

Mots-clés

Indicateurs de l'état de santé, enquêtes sur la santé, transition entre états de santé, modèles logistiques, études longitudinales, modèle de croissance multiniveaux.

Auteures

Julie Bernier (1-613-951-4556; julie.bernier@statcan.gc.ca) et Keiko Asakawa travaillent à la Division de l'analyse de la santé et Yan Feng, à la Division de la statistique du revenu, à Statistique Canada, Ottawa (Ontario) K1A 0T6.

Les données longitudinales de l'Enquête nationale sur la santé de la population (ENSP) réalisée par Statistique Canada peuvent être utilisées pour évaluer la dynamique de l'état de santé. Pendant plus d'une décennie, des données répétées ont été recueillies tous les deux ans auprès des membres des échantillons de l'ENSP. Le calcul d'estimations d'après des données issues de mesures répétées est souvent plus facile si l'on opte pour une approche fondée sur un modèle (multiniveaux) de courbe de croissance[1] en vue d'estimer les variations intra-individu (niveau 1) et inter-individus (niveau 2). Un tel modèle permet de représenter la dynamique des résultats par une trajectoire et d'examiner les associations entre les caractéristiques socioéconomiques et les déterminants de la santé, d'une part, et les trajectoires de la qualité de vie liée à la santé (QVLS), d'autre part.

Comme dans toute méthode de régression, l'utilité des estimations dépend de la mesure dans laquelle les hypothèses de modélisation sont satisfaites. Dans les modèles à un seul niveau, les hypothèses au sujet des spécifications du modèle (p. ex. linéarité, variables omises, interactions) et des spécifications stochastiques (p. ex. hétéroscédasticité, normalité des erreurs) doivent être évaluées minutieusement[2,3]. Le présent article porte sur l'hypothèse de normalité des erreurs dans les conditions d'un modèle de courbe de croissance. Dans un tel modèle, où les enquêtés sont considérés comme étant les unités de niveau 2 et les cycles (périodes) d'enquête chez

chaque enquêté comme étant les unités de niveau 1, l'hypothèse de normalité des erreurs signifie la normalité univariée des résidus au niveau 1 et la normalité univariée ou multivariée (si plus d'un paramètre est considéré comme étant aléatoire) des composantes aléatoires au niveau 2. Le fait que l'hypothèse de normalité ne soit pas vérifiée au niveau 1 ne causera pas de biais dans l'estimation des effets fixes, mais elle en introduira un dans les erreurs-types aux deux niveaux, influençant ainsi la validité des intervalles de confiance et des tests d'hypothèse. L'estimation des effets fixes de niveau 2 ne sera pas biaisée par la non-normalité des erreurs au niveau 2. Cependant, l'existence d'une asymétrie aura une incidence sur les inférences au niveau 2[2].

Dans les analyses des données longitudinales provenant d'enquêtes telles que l'ENSP, il faut considérer l'hypothèse de normalité des erreurs, car les résultats en matière de santé de la population, tels que la QVLS, ont souvent une distribution asymétrique. Le *Health Utilities Index Mark 3* (HUI3) est une mesure de la QVLS recueillie dans le cadre de l'ENSP. Il s'agit d'un indicateur générique, à attributs multiples, fondé sur les préférences mesurées sur une échelle continue, qui décrit l'état de santé au moyen d'une mesure sommaire unique variant de -0,36 à 1,00 (1,00 = santé parfaite; 0,00 = décès; -0,36 = état pire que le décès)[4]. Toutefois, étant donné la distribution fortement asymétrique de l'HUI3[5,6], l'hypothèse de normalité des erreurs peut être violée quand elle est utilisée pour estimer des modèles de courbe de croissance. Si l'on introduit les déterminants de la santé comme variables dans le modèle, les erreurs-types des paramètres ne peuvent pas être estimées correctement.

Un moyen de traiter cette violation de l'hypothèse de normalité consiste à transformer la variable de résultat de façon à améliorer la distribution des erreurs. (La transformation a pour but de produire des résidus, et non des variables dépendantes, dont la distribution n'est pas asymétrique.) Dans la présente étude, nous évaluons l'utilité de la transformation *arcsinus*, qui stabilise la variance et améliore la symétrie des résidus[7]. Une étude antérieure a montré qu'en utilisant une variable HUI3 non transformée comme variable de résultat, les scores HUI3 prédits se situaient en dessous de la limite inférieure théorique de -0,36[1].

Un examen préliminaire de la transformation arcsinus de la forme (arcsin[2 × (HUI3 + 0,36) / (1 + 0,36) -1]) a produit des scores HUI3 rétrotransformés se situant au-dessus de la borne inférieure théorique de -0,36. Bien que la transformation arcsinus ait été utilisée auparavant[8], autant que nous sachions, elle n'a pas été appliquée à l'analyse de données longitudinales sur la santé de la population. Donc, on ignore comment l'hypothèse de normalité des erreurs sous un modèle de courbe de croissance est traitée dans cette transformation.

L'objectif principal de la présente étude était d'évaluer la faisabilité de l'utilisation de la transformation arcsinus issue de la famille de fonctions trigonométriques pour estimer les modèles de courbe de croissance lorsque la distribution des résidus ne suit pas une loi normale. L'évaluation a été effectuée pour un modèle socioéconomique simple contenant comme variables l'état matrimonial, le niveau de scolarité et le revenu du ménage. Deux autres transformations ont également été examinées, l'une issue d'une famille de fonctions logarithmiques (transformation logarithme naturel) et l'autre, d'une famille de fonctions exponentielles (transformation racine carrée). La performance de ces trois transformations a été comparée.

Une autre difficulté que pose la transformation du modèle est l'interprétation des résultats d'estimation. Comme la rétrotransformation des coefficients estimés est difficile, voire impossible, l'objectif secondaire de l'étude était de présenter une approche graphique pour interpréter les résultats d'estimation en se basant sur un modèle dont la variable dépendante est transformée.

Une étude de cas portant sur la performance des modèles de courbe de croissance obtenus en procédant à divers types de transformation de la variable de résultat HUI3 a été réalisée. Le but était de proposer une approche pragmatique du traitement de l'hypothèse de normalité des erreurs et non de trouver le modèle le plus adéquat parmi les types possibles de formes fonctionnelles, de techniques d'estimation ou de spécifications du modèle. La présente étude montre les possibilités qu'offre l'application de la transformation arcsinus à des modèles de courbe de croissance au moyen d'une comparaison des résultats produits par cette transformation à ceux d'autres méthodes de transformation utilisées fréquemment.

Méthodes

Source des données

Les données proviennent de la composante des ménages des cycles de 1994-1995 à 2006-2007 de l'Enquête nationale sur la santé de la population (ENSP), qui était conçue pour recueillir des renseignements longitudinaux sur l'état de santé et les caractéristiques sociodémographiques de la population canadienne. La population cible comprenait les membres des ménages des dix provinces en 1994-1995, sauf les résidents des réserves indiennes et des terres de la Couronne, des établissements de soins de santé et de certaines régions éloignées de l'Ontario et du Québec, les membres à temps plein des Forces canadiennes, ainsi que tous les résidents (militaires et civils) des bases des Forces canadiennes.

Pour étudier les trajectoires des scores HUI3, nous nous sommes servis des données provenant du fichier carré longitudinal de l'ENSP. Ce fichier carré contient les données sur les 17 276 participants au cycle 1 (1994-1995), indépendamment de leur profil de réponse aux six cycles suivants. La taille de l'échantillon longitudinal est restée la même pour tous les cycles[9,10].

La présente étude porte sur les personnes âgées de 40 à 99 ans en 1994-1995 qui ont répondu aux questions de l'HUI3. L'exclusion du petit nombre de personnes de 100 ans et plus n'a pas eu d'incidence sur les estimations des paramètres dans le modèle de régression. La population cible comprenait les 252 enquêtés qui ont été placés en établissement à un moment donné au cours des six cycles de suivi. Les 1 295 personnes qui sont décédées durant le suivi ont également été incluses, mais uniquement pour le cycle durant lequel leur décès a été déclaré; pour les cycles subséquents, l'information a été catégorisée comme manquante. L'échantillon final comprenait 7 784 personnes.

Variable de résultat

La variable de résultat est le score HUI3, une variable continue dont la valeur varie de -0,36 à 1,00 (1,00 = santé parfaite; 0,00 = décès; -0,36 = état pire que le décès). Afin de comparer les modèles sous diverses transformations, les scores HUI3 ont été transformés comme il est décrit plus loin.

Variables indépendantes

Les formes linéaire et non linéaire d'une variable indiquant l'âge des enquêtés ont été incluses dans le modèle pour représenter la dimension temporelle dans les analyses. La variable d'âge a été centrée à 57 ans, qui était l'âge moyen à la période de référence. Le sexe, l'état matrimonial, le niveau de scolarité et le revenu du ménage ont été ajoutés comme variables indépendantes. Le sexe (féminin comme groupe de référence) a été inclus sous forme de variable temporellement invariante. L'état matrimonial, le niveau de scolarité et le revenu du ménage ont été inclus sous forme de covariables variant temporellement. L'état matrimonial a été réparti en deux catégories, à savoir marié(e)/union de fait/vit avec un(e) partenaire (groupe de référence) ou célibataire/séparé(e)/divorcé(e)/veuf(ve). Les catégories de niveau de scolarité étaient pas de diplôme d'études secondaires ou au moins un diplôme d'études secondaires (groupe de référence). Le revenu du ménage a été catégorisé comme étant faible (moins de 15 000 $), moyen (de 15 000 $ à 29 999 $) ou élevé (30 000 $ ou plus; groupe de référence).

Deux jeux de variables indicatrices variant temporellement ont été créés comme variables de contrôle : lieu de résidence (1 si placement en établissement, 0 autrement) et l'état de décès (1 si décédé, 0 autrement). Pour tenir compte des effets de la mortalité dans les analyses[1], l'enregistrement de la première déclaration du décès a été retenu dans les analyses en attribuant une valeur de HUI3 = 0,00 à la variable dépendante. Pour chacune des variables indépendantes, la dernière valeur observée a été attribuée au premier enregistrement du décès; dans les enregistrements subséquents, les valeurs ont été catégorisées comme manquantes.

Modélisation

Les données de l'ENSP sont des mesures répétées des caractéristiques des enquêtés au cours des six cycles de collecte des données. La structure hiérarchique des données – mesures répétées emboîtées dans les enquêtés – peut être modélisée en utilisant un modèle de croissance à deux niveaux. Un modèle de croissance multiniveaux intègre simultanément les variations intra-personne et inter-personnes. Le modèle intra-sujet (niveau 1) a été spécifié sous forme d'une fonction d'un ensemble de paramètres de croissance et de caractéristiques individuelles fixes associées à une erreur de mesure. Les paramètres de croissance et les caractéristiques individuelles variant temporellement, qui sont une source d'hétérogénéité, ont été spécifiés dans le modèle inter-sujets (niveau 2) pour refléter la variation de la croissance au sein de la population.

Comme dans une étude antérieure[1], l'âge a été exprimé sous une forme cubique, à savoir un taux de croissance linéaire (Âge), un taux de croissance quadratique ($Âge^2$) et un taux de croissance cubique ($Âge^3$). Seule l'ordonnée à l'origine a été considérée comme variant aléatoirement d'une personne à l'autre, ce qui représente une forme plus simple que celle présentée dans l'étude précédente[1].

Le modèle de croissance contenant le revenu du ménage, le niveau de scolarité, l'état matrimonial, le décès et le placement en établissement au niveau 1, et le sexe au niveau 2, est :

Niveau 1 :

$$Y_{ij} = \beta'_{0j} + \beta'_{1j}\,\hat{A}ge_{ij} + \beta'_{2j}\,\hat{A}ge^2_{ij} + \beta'_{3j}\,\hat{A}ge^3_{ij} + \sum_{n=4}^{N}\beta'_{nj}X_{nij} + \varepsilon'_{ij}$$

$$\varepsilon'_{ij} \sim N(0, \sigma'^2_{\varepsilon})$$

Niveau 2 :

$$\beta'_{0j} = \beta'_{00} + \beta'_{10}\,\text{Sexe}_j + \zeta'_{0j}$$

$$\zeta'_{0j} \sim N(0, \sigma'^2_0)$$

où Y_{ij} représente le score HUI3 de l'individu j au cycle i. Les quatre coefficients β du niveau 1 sont les vrais paramètres de croissance qui varient d'un individu à l'autre. σ^2_{ε} décrit l'écart aléatoire intra-individu du score HUI3 par rapport à sa propre trajectoire. β_{00} représente le score HUI3 moyen de la population à l'âge de 57 ans et σ^2_0 représente l'écart aléatoire des scores HUI3 moyens entre individus (à l'âge de 57 ans)[3].

Un ensemble de données personne-période a été créé pour l'analyse des modèles de croissance multiniveaux. La date de naissance et la date de l'entrevue enregistrées dans le fichier de microdonnées de l'ENSP ont permis d'utiliser les données de l'enquête temporellement déstructurées (les âges réels des enquêtés ne varient pas nécessairement d'un intervalle de deux ans entre les périodes évaluées) grâce au calcul de l'âge réel des enquêtés[3]. Par conséquent, le modèle a été ajusté en utilisant les valeurs numériques réelles de la variable d'*âge* (la différence entre la date de l'entrevue et la date de naissance autodéclarée) comme variable temporelle, ce qui a rendu les données personne-période temporellement déstructurées. Comme dans la plupart des études longitudinales,

les données étaient déséquilibrées en raison de l'érosion de l'échantillon. Les modèles de courbe de croissance permettent de calculer des estimations en se servant de données temporellement déstructurées et déséquilibrées. Des 7 784 personnes que contenait l'échantillon, 2 989 (38,4 %) possédaient six enregistrements, 1 546 (19,9 %) en possédaient cinq, 1 141 (14,7 %) en possédaient quatre, 928 (11,9 %) en possédaient trois, 748 (9,6 %) en possédaient deux et 432 (5,6 %) n'en possédaient qu'un seul.

Transformation

La transformation arcsinus a été utilisée pour transformer la variable dépendante (HUI3) afin d'accroître la normalité des résidus. La transformation arcsinus est habituellement appliquée à une variable ayant une étendue de [-1, 1]. Si une variable X varie de -1 à 1, arcsin(X) variera de moins l'infini à plus l'infini. Cependant, la variable HUI3 est bornée par les valeurs -0,36 et 1,00. D'où la transformation arcsinus ne peut pas être appliquée efficacement pour modifier la distribution des scores HUI3, à cause de la borne inférieure théorique de -0,36. Pour faciliter la transformation arcsinus, nous avons d'abord transformé linéairement la variable HUI3 de manière que les scores HUI3 transformés soient bornés par les valeurs -1 et +1 en nous servant de l'équation suivante :

$$2 \times \frac{HUI3 + 0{,}36}{1 + 0{,}36} - 1$$

La transformation arcsinus a été appliquée en utilisant les scores HUI3 transformés (arcsin [2 × (HUI3 + 0,36) / (1 + 0,36) -1]). Cette transformation a amélioré la prédiction de la trajectoire pour une population vieillissante en permettant que les scores HUI3 prédits rétrotransformés soient à l'intérieur de la borne théorique inférieure de -0,36. Par contre, dans l'étude antérieure, les scores HUI3 prédits tombaient au delà de -0,36 quand une variable HUI3 non transformée était utilisée comme variable de résultat[1].

Évaluation de l'hypothèse de normalité

L'hypothèse de normalité des erreurs a été évaluée en comparant les statistiques d'asymétrie entre les divers modèles. Les statistiques de plus haut niveau, comme le kurtosis, n'ont pas été estimées. Dans les analyses, nous avons considéré que les erreurs suivaient presque une loi normale si la statistique d'asymétrie était nulle[11]. Nous avons considéré qu'une amélioration du respect de l'hypothèse de normalité des erreurs avait eu lieu si les statistiques d'asymétrie des erreurs de niveau 1 et de niveau 2 d'un modèle particulier étaient plus proches de 0 que celles d'un autre modèle. Nous avons également utilisé des graphiques gausso-arithmétiques (*normal probability plots*) pour évaluer la distribution des résidus à chacun des deux niveaux. L'obtention d'une droite lorsque l'on représente graphiquement les scores normaux théoriques en fonction des résidus standardisés signifie que les résidus suivent une loi normale[12].

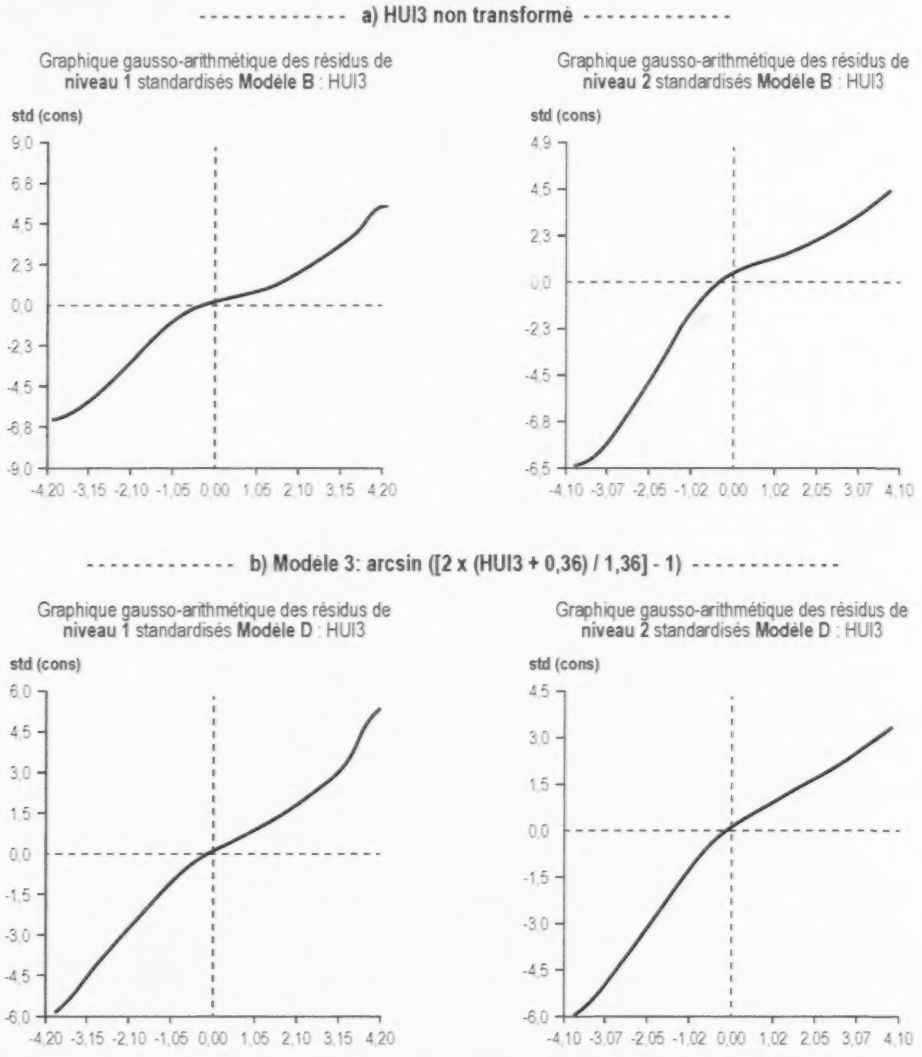
Pour déterminer si la transformation arcsinus était appropriée, nous avons comparé la distribution des résidus du modèle transformé à celles des résidus du modèle non transformé et de deux autres modèles, fondés chacun sur une forme de transformation différente de la variable dépendante, à savoir le logarithme naturel[13] et la racine carrée[7]. Nous avons évalué l'hypothèse de normalité des erreurs en comparant les statistiques d'asymétrie et les graphiques gausso-arithmétiques obtenus pour les quatre modèles.

Tableau 1
Comparaison de l'ajustement de diverses trajectoires de changement polynomiales

	HUI3 non transformé	Modèle 1	Modèle 2	Modèle 3
Effets fixes				
Ordonnée à l'origine	0,882576**	0,627531**	0,952846**	1,076387**
Âge moyen	-0,001724**	-0,000962**	-0,000682**	-0,005272**
Âge moyen au carré	-0,000050**	-0,000033**	-0,000023**	-0,000018**
Âge moyen au cube	-0,000004**	-0,000003**	-0,000002**	-0,000008**
Décédé(e)	-0,668605**	-0,489359**	-0,340475**	-1,134217**
Placé(e) en établissement	-0,459094**	-0,352409**	-0,253395**	-0,810786**
Femme†				
Homme	0,0006329	0,003433	0,002511	0,023969*
Diplôme d'études secondaires ou plus				
Pas de diplôme d'études secondaires†	-0,037715**	-0,023636**	-0,016734**	-0,082577**
Marié(e)/union de fait/vit avec un(e) partenaire†				
Célibataire	-0,01384	-0,007427	-0,005187	-0,03341
Divorcé(e)/séparé(e)/veuf(ve)	-0,014254**	-0,00806**	-0,005672**	-0,030819**
Revenu élevé				
Revenu moyen	-0,020276**	-0,013019**	-0,009237**	-0,042748**
Revenu faible	-0,045075**	-0,029279**	-0,020741**	-0,091397**
Effets aléatoires				
Niveau 1 (σ_{ε}^2 (intra-personne))	0,024419**	0,010790**	0,00542**	0,106314**
Niveau 2 ($\sigma_{\zeta 0}^2$ (ordonnée à l'origine))	0,016104**	0,006713**	0,003357**	0,069918**
Asymétrie				
Niveau 1	-1,35	-1,77	-1,88	-0,74
Niveau 2	-1,59	-1,91	-1,93	-1,06

* valeur significativement différente de l'estimation pour le groupe de référence ($p < 0,0005$)
** valeur significativement différente de l'estimation pour le groupe de référence ($p < 0,0001$)
† catégorie de référence
... n'ayant pas lieu de figurer
Nota : HUI3 non transformé comme variable dépendante
Modèle 1 : ln (HUI3 + 1) comme variable dépendante
Modèle 2 : racine carrée de [(HUI3 + 0,36) / 1,36] comme variable dépendante
Modèle 3 : arcsin[2 × (HUI3 + 0,36) / (1 + 0,36) –1] comme variable dépendante
Source : Enquête nationale sur la santé de la population, 1994-1995 à 2006-2007, fichier carré longitudinal

Figure 1
Comparaison des graphiques gausso-arithmétiques des résidus



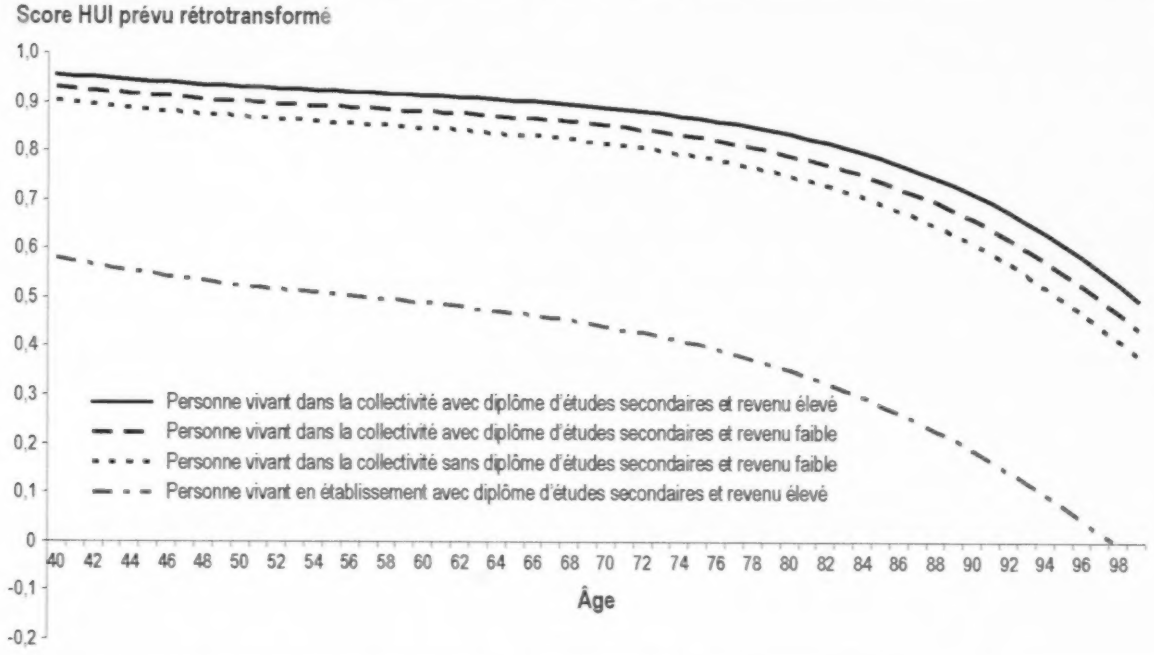
Source : Enquête nationale sur la santé de la population, 1994-1995 à 2006-2007, fichier carré longitudinal

Interprétation des modèles

En partant des résultats du modèle arcsinus, nous avons construit un graphique représentant les trajectoires estimées en fixant les valeurs des variables explicatives à divers niveaux. En particulier, nous avons tracé les trajectoires des scores HUI3 rétrotransformés en fonction du sexe, de l'état matrimonial, du niveau de scolarité, du revenu du ménage et du lieu de résidence (collectivité ou établissement).

Toutes les analyses ont été effectuées au moyen des logiciels SAS et MLwiN. Les modèles ont été pondérés en utilisant les poids de sondage afin de tenir compte des probabilités de sélection inégales des participants à l'ENSP, et les poids ont été appliqués au deuxième niveau du modèle. Les estimations de la variance n'ont pas été corrigées pour tenir compte du plan de sondage complexe de l'enquête.

Figure 2
Trajectoire prévue du score HUI3 pour quatre groupes de femmes mariées de 40 ans et plus, selon l'année d'âge, le niveau de scolarité et le revenu du ménage, Canada, territoires non compris, 1994-1995 à 2006-2007



Source : Enquête nationale sur la santé de la population, 1994-1995 à 2006-2007, fichier carré longitudinal

Résultats

Étude descriptive préliminaire des trajectoires de l'état de santé

Afin de trouver une forme fonctionnelle appropriée pour le sous-modèle de niveau 1 et pour résumer l'évolution de l'état de santé au cours du temps, nous avons examiné des courbes de croissance empiriques du score HUI3 avec des trajectoires non paramétriques lisses selon le groupe d'âge. La trajectoire du score HUI3 moyen selon l'âge donnait à penser que le sous-modèle de niveau 1 n'était pas linéaire[1]. La trajectoire de la variable HUI3 comporte une trajectoire quadratique et une trajectoire cubique (graphiques non présentés). Au lieu de choisir une forme polynomiale unique pour chaque personne, nous avons sélectionné le polynôme d'ordre le plus élevé pour résumer la variation individuelle pour toute personne. Par conséquent, nous avons inclus des termes linéaire, quadratique et cubique de l'Âge dans tous les modèles.

Résultats des régressions

Étant donné la complexité de la rétrotransformation des coefficients estimés dans le modèle arcsinus de l'HUI3 pour retourner à l'échelle originale, nous n'interprétons pas séparément les paramètres estimés. En outre, quand une transformation est appliquée à la variable de résultat, les paramètres estimés du modèle transformé ont les propriétés des moindres carrés se rapportant aux observations transformées seulement et non aux observations originales[14]. Par conséquent, nous ne présentons que les paramètres des modèles transformés dans la présente étude.

Comparaison des modèles proposés

Les comparaisons entre les modèles révèlent que, pour le modèle avec transformation arcsinus (modèle 3), l'asymétrie des résidus de niveau 1 est de -0,74, valeur qui, de celles produites par tous les modèles, est la plus proche de 0 (tableau 1). En particulier, pour le modèle à transformation logarithmique (modèle 1) et celui à transformation racine carrée (modèle 2), la statistique d'asymétrie des résidus de niveau 1 atteint -1,77 et -1,88, respectivement. La comparaison des graphiques gausso-arithmétiques pour les scores HUI3 non transformés (figure 1a) avec ceux obtenus par transformation arcsinus montre que le graphique des résidus pour le modèle arcsinus (modèle 3, figure 1b) semble être celui qui s'approche le plus de la linéarité parmi les quatre modèles (figures pour les modèles 1 et 2 non présentées).

La statistique d'asymétrie des résidus de niveau 2 est environ de -1,90 dans le modèle 1 ainsi que dans le modèle 2, encore plus éloignée de 0 que pour le niveau 1. Pour la transformation arcsinus (modèle 3), la statistique d'asymétrie est de -1,06, c'est-à-dire la valeur la plus proche de 0 parmi les modèles étudiés. Les graphiques gausso-arithmétiques des résidus standardisés au niveau 2 montrent

que le tracé obtenu pour le modèle 3 est celui qui s'approche le plus de la linéarité parmi les quatre modèles, quoique les queues s'écartent encore de la normalité à l'extrémité supérieure de la distribution (figure 1b).

Interprétation des modèles : une approche graphique

À titre d'exemple, nous présentons les trajectoires de l'état de santé de femmes ayant certains profils sociodémographiques fondées sur le modèle avec transformation arcsinus (modèle 3) (figure 2). Quand toutes les autres caractéristiques sociodémographiques sont maintenues constantes, les trajectoires d'état de santé des femmes vivant dans la collectivité diffèrent considérablement de celles des femmes vivant en établissement. Les femmes qui vivent dans la collectivité sont, en moyenne, en nettement meilleure santé, et la différence s'accentue avec l'âge. Le niveau de scolarité est également associé aux trajectoires de l'état de santé. Chez les femmes vivant dans la collectivité dans un ménage à faible revenu, la trajectoire du score HUI3 de celles n'ayant pas obtenu de diplôme d'études secondaires est inférieure à la trajectoire de celles ayant au moins décroché leur diplôme d'études secondaires. Le revenu du ménage est également corrélé à la variation des trajectoires de l'état de santé.

Discussion

Lorsque l'on évalue l'état de santé d'une population, il n'est pas rare que les mesures continues de la QVLS ne suivent pas une loi normale, si bien que les hypothèses de normalité sont violées dans les analyses de régression. Le problème touche l'estimation des modèles de courbe de croissance en ce sens que la violation de l'hypothèse de normalité produit un biais dans les erreurs-types, ce qui influence la validité des intervalles de confiance et des tests d'hypothèse. La présente analyse s'appuie sur une étude de cas pour déterminer si la transformation de la variable de résultat au moyen de la formule arcsin [2 × (HUI3 + 0,36) / (1 + 0,36) - 1] permettrait de résoudre ce problème. Les résultats montrent que le modèle avec transformation arcsinus rend la distribution des résidus moins asymétrique. La symétrie de la distribution des résidus est améliorée considérablement comparativement aux distributions produites par le modèle non transformé et par les modèles avec transformation par le logarithme naturel et par la racine carrée. Une approche graphique, consistant à tracer les trajectoires des scores HUI3 prédits rétrotransformés est également présentée, tout en reconnaissant que l'interprétation des résultats d'estimation fondés sur un modèle avec variable dépendante transformée non linéairement est complexe.

Autant que nous sachions, l'étude de cas est unique en ce sens que la transformation arcsinus n'a pas été appliquée à un modèle de courbe de croissance dans l'analyse des données d'enquête sur la santé de la population et que les résultats de la transformation sont meilleurs que ceux du modèle non transformé ou des modèles avec transformation par logarithme naturel ou par racine carrée. L'approche graphique est un moyen simple d'interpréter les résultats du modèle arcsinus en évitant les complications que comporte l'exploration de la rétrotransformation de l'ensemble de coefficients estimés.

Plusieurs points méritent d'être soulignés. Premièrement, nous avons attribué à l'enregistrement du décès d'un enquêté les dernières valeurs observées des variables explicatives. Cette approche simple n'est peut-être pas optimale, car d'autres méthodes sont disponibles[15]. Cependant, moins de 1 % des enregistrements correspondant au premier enregistrement de décès utilisés dans les analyses ont été imputés; les valeurs pour les cycles subséquents ont été catégorisées comme manquantes, de sorte que tout biais résultant de cette approche est vraisemblablement minime. Deuxièmement, les trois modèles testés ne constituent pas une représentation exhaustive des transformations possibles. Troisièmement, certaines mesures de la QVLS, y compris le score HUI3, sont sujettes à un effet de plafonnement. Cependant, cet effet n'est vraisemblablement pas un problème dans la présente étude, parce que 10 % seulement des enregistrements obtenus pour les enquêtés de 40 ans et plus (moins de 1 % de ceux pour les enquêtés de 65 ans et plus) contenaient un score HUI3 de 1,00 (santé parfaite). Quatrièmement, les variables explicatives comprises dans les modèles ont été choisies uniquement pour illustrer l'utilité de la transformation du score HUI3. Avant de choisir un modèle définitif, une plus grande attention doit être accordée aux choix des covariables et à l'examen d'un effet de pente aléatoire éventuel. Cinquièmement, en théorie, les scores d'utilité numériques tels que l'HUI3 sont uniques jusqu'à une transformation linéaire positive[16], mais la transformation arcsinus n'est pas linéaire. Néanmoins, la méthode proposée est pragmatique et utile, car les courbes des scores HUI3 rétrotransformés permettent de visualiser plus facilement d'importantes hétérogénéités dans les trajectoires de l'état de santé.

L'étude de cas a montré que la transformation arcsinus est un moyen pratique de traiter l'hypothèse de normalité des erreurs dans la modélisation multiniveaux. Elle est également utile pour estimer l'effet de divers déterminants de la santé sur les trajectoires de l'état de santé. La méthode peut être appliquée en se servant de la plupart des progiciels statistiques et peut décrire les variations des trajectoires de l'état de santé entre divers groupes socioéconomiques. L'approche peut également être utilisée pour évaluer d'autres déterminants de la santé. La présente étude de cas est un premier effort en vue d'introduire cette transformation dans un contexte de modélisation multiniveaux. Une étude plus approfondie devra être effectuée afin d'examiner les possibilités de la transformation arcsinus pour d'autres types de transformations, méthodes d'estimation et populations. ■

Références

1. H. Orpana, N. Ross, D. Feeny *et al.*, « Histoire naturelle de la qualité de vie liée à la santé : une étude de cohorte de dix ans », *Rapports sur la santé*, 20(1), 2009, p. 31-38.
2. S.W. Raudenbush et A.S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods*, Second Editon, Thousand Oaks, California, Sage Publications, 2002, p. 266-267.
3. J.D. Singer et J.B. Willett, *Applied Longitudinal Data Analysis: Modelling Change and Event Occurrence*, Toronto, Oxford University Press, 2003.
4. D. Feeny, W. Furlong, G.W. Torrance *et al.*, « Multi-attribute and single-attribute utility functions for the Health Utilities Index Mark 3 System », *Medical Care*, 40(2), 2002, p. 113-128.
5. K. Eng et D. Feeny, « Comparing the health of low income and less well educated groups in the United States and Canada », *Population Health Metrics*, 5(10), 2007, p. 103.
6. E. van Doorslaer et A.M. Jones, « Inequalities in self-reported health: Validation of a new approach to measurement », *Journal of Health Economics*, 22, 2003, p. 61-87.
7. D. Kleinbaum, L. Kupper, K. Muller et A. Nizam, *Applied Regression Analysis and Other Multivariable Methods*, Third edition, Toronto, Duxbury Press, 1997, p. 252.
8. A. Hertley, K.F. Ho, C. McConkey et J.I. Geh, « Pathological complete response following pre-operative chemoradiotherapy in rectal cancer: analysis of phase II III trials », *The British Journal of Radiology*, 78, 2005, p. 934-938.
9. L. Swain, G. Catlin et M.P. Beaudet, « Enquête nationale sur la santé de la population – une enquête longitudinale », *Rapports sur la santé*, 10(4), 1999, p. 73-89.
10. Statistique Canada, *Enquête nationale sur la santé de la population, Cycles 1 à 7 (1994-1995 à 2006-2007), volet ménages, documentation longitudinale*, Ottawa, Statistique Canada, 2007.
11. P.J. Bickel et K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Oakland, California, Holden-Day, 1977, p. 388.
12. R.A. Johnson et G.K. Bhattacharyya, *Statistics, Principles and Methods*, Toronto, John Wiley & Sons Inc., 1992.
13. W.G. Manning et J. Mullahy, « Estimating log models: Transform or not to transform? », *Journal of Health Economics*, 20(4), 2001, p. 461-494.
14. J. Neter, W. Wasserman et M.H. Kutner, *Applied Linear Statistical Models*, Second Edition, Homewood, Illinois, Richard P. Irwin, Inc., 1985, p. 141.
15. R.J.A. Little et D.B. Rubin, *Statistical Analysis with Missing Data*, Second Edition, Hoboken, New Jersey, Wiley, 2002.
16. G.A. Jehle et P.J. Reny, *Advanced Microeconomic Theory*, Reading, Massachusetts, Addison-Wesley, 1998.